

Self-Organizing Maps. An application to the OGLE data and the Gaia Science Alerts

Łukasz Wyrzykowski^{*,†} and Vasily Belokurov^{*}

^{*}*Institute of Astronomy, University of Cambridge, UK*

[†]*Warsaw University Astronomical Observatory, Poland*

Abstract.

Self-Organizing Map (SOM) is a promising tool for exploring large multi-dimensional data sets. It is quick and convenient to train in an unsupervised fashion and, as an outcome, it produces natural clusters of data patterns. An example of application of SOM to the new OGLE-III data set is presented along with some preliminary results.

Once tested on OGLE data, the SOM technique will also be implemented within the Gaia mission's photometry and spectrometry analysis, in particular, in so-called classification-based Science Alerts. SOM will be used as a basis of this system as the changes in brightness and spectral behaviour of a star can be easily and quickly traced on a map trained in advance with simulated and/or real data from other surveys.

Keywords: Observations: photometry, spectrometry; Astronomical catalogues; Stars: variable and peculiar; Neural networks

PACS: 95.75.De, 95.75.Fg, 95.80.+p, 97.30.-b, 84.35.+i

SOM - WHAT IS IT?

The Self-Organizing Map is described by its author Teuvo Kohonen ([1]) as a map reflecting topological ordering. It is a list of weight vectors organised as a 2D grid of map nodes (neurones). Each datum is mapped onto a node associated with the nearest weight vector, e.g. the one with the smallest Euclidean distance from the data pattern, but any kind of similarity measure can be used. The SOM organises itself during a competitive and unsupervised learning process. Each pattern is shown to the SOM (randomly or sequentially) and the closest node ("winner") is found. Then all the neighbouring nodes of the winner are adjusted with the learning rate α :

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(\mathbf{x}(t) - \mathbf{m}_i(t)) \quad (1)$$

In the next step, the neighbourhood radius and the learning rate is decreased and next pattern is shown. The process continues until the set of patterns is exhausted or the learning rate reaches 0.

Describing a pattern

This is the crucial and the most challenging step when working with SOMs. The pattern should be described in the most efficient way, however, with emphasis on characteristic features of the classes. The advantage of SOMs is that patterns can be as long

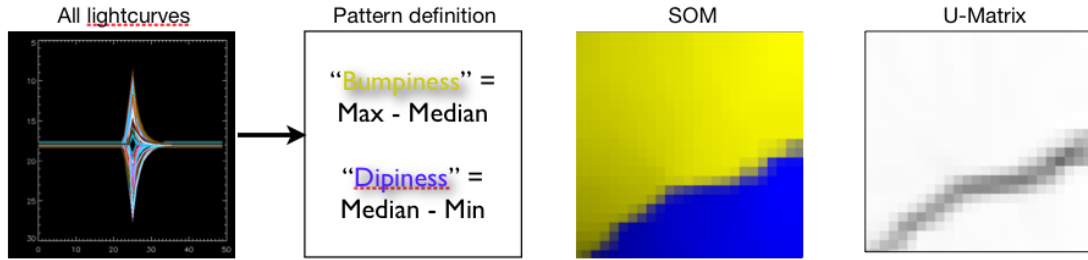


FIGURE 1. The simple example of SOM. Two kinds of light curves - a bump and a dip with different amplitudes - are described by two parameters, therefore each pattern contains two values: "bumpiness" and "dipiness". Resulting SOM is visualised by assigning its vectors to the RGB channels (here Red and Green channels were joined). U-Matrix plot shows the distribution of distances between the nodes of the SOM. This SOM easily disentangles the two light curve classes.

as we like. For example, a pattern can contain a whole picture, a binned light curve or a spectrum, a vector of statistical parameters, binned periodogram, or the combination of these.

OPTICAL GRAVITATIONAL LENSING EXPERIMENT

OGLE has run since 1992 at the University of Warsaw (Poland). It uses a dedicated 1.3m telescope in Chile, which continuously monitors hundreds of millions of stars towards

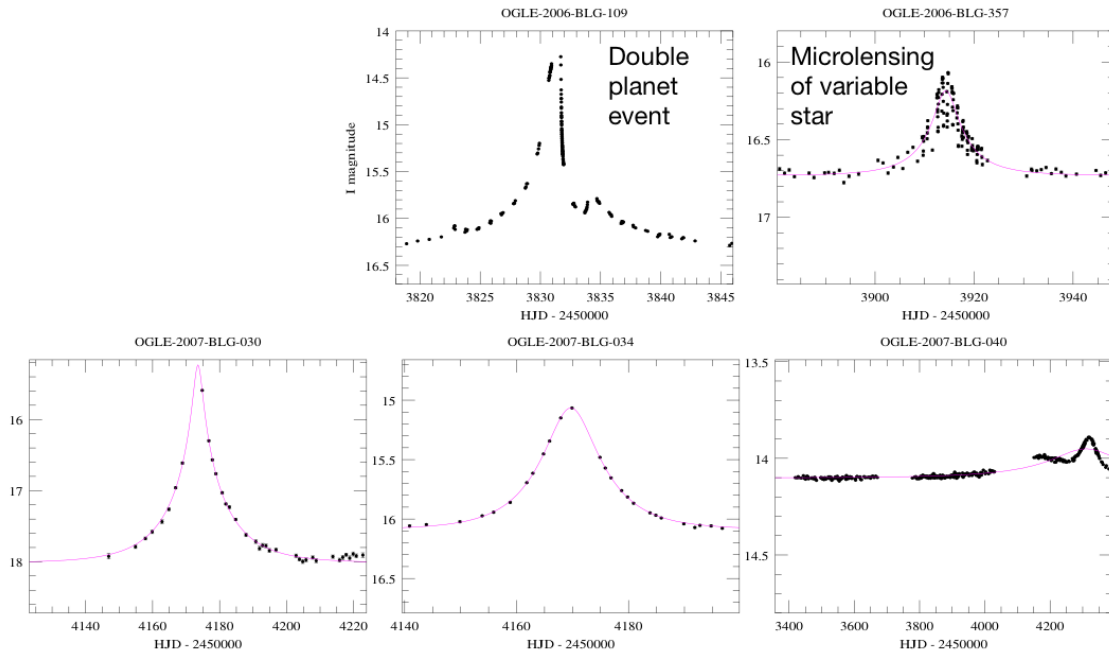


FIGURE 2. Examples of OGLE-III microlensing events with standard model curve (magenta line).

the Galactic Centre and Magellanic Clouds in order to detect gravitational microlensing events ([2]). As a natural by-product of such search, vast number of variable stars are being discovered and monitored. Presently, nearly a billion objects towards the Galactic bulge and Magellanic Clouds need to be investigated and classified into variability classes¹.

In early 2009 OGLE will upgrade its camera to 34 CCDs covering about 1 sq. deg. in one exposure. It will create vast data sets which will have to be analysed in an automated manner.

OGLE's Early Warning System (EWS) [3] is detecting on-going microlensing event almost in real-time. Since 2002 it has detected nearly 4000 candidates for events. Most of them are following typical Paczynski's curve (see Fig. 2), but there are also anomalous events due to e.g. binary lens, parallax effect or presence of a planet (e.g. [4]). These anomalous light curves vary enormously in their shape and present a challenge for a quick and robust classification.

The SOM shown in Figure 3 was trained on patterns containing three microlensing

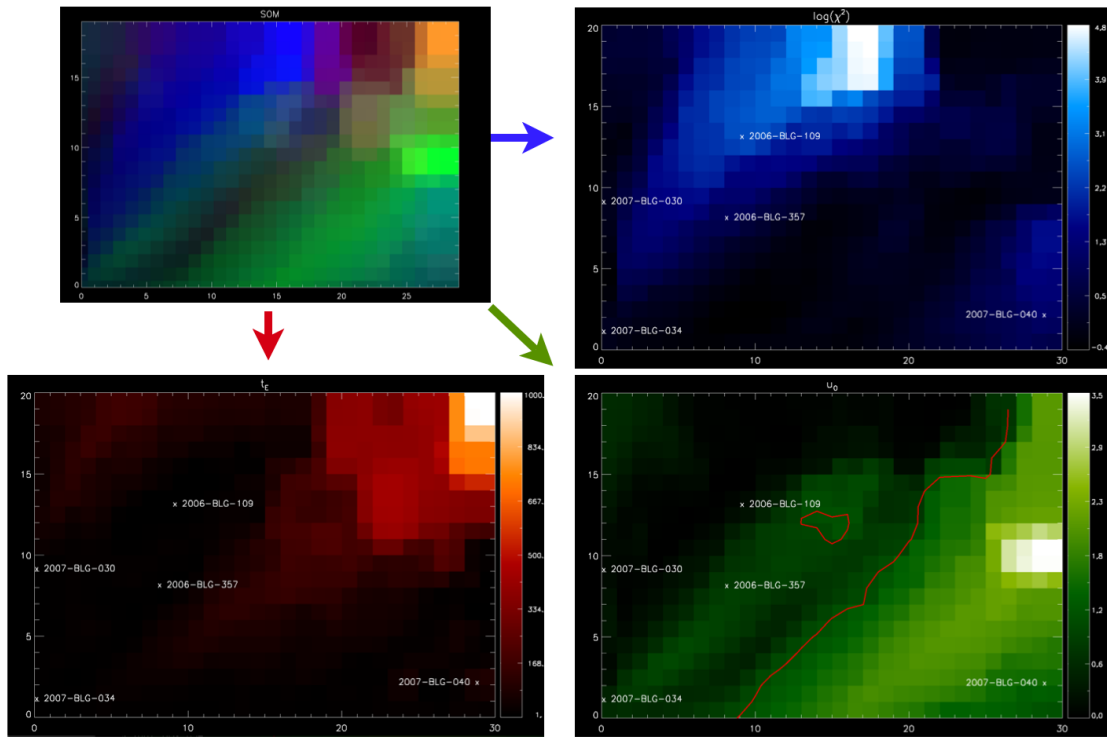


FIGURE 3. SOM trained on OGLE-III microlensing events parameters (t_E , u_0 and χ^2). The SOM (upper left) can be also visualised by splitting each RGB channel separately for each parameter. Positions of exemplary microlensing events (see 2) are marked. Boundary of events with lens crossing the Einstein Radius ($u_0 < 1$) is marked as a red contour.

¹ For the first attempt of applying SOM to the variable stars detected by OGLE see http://www.ast.cam.ac.uk/~vasily/ogle_som/

model parameters: time-scale (t_E), impact parameter (u_0) and goodness of the model fit ($\log\chi^2$). Every parameter was coded in different RGB channel and three maps show each channel separately. Positions of exemplary events are marked on the maps. Anomalous and spurious events have high χ^2 (e.g. 2006-BLG-109). Such trained map can be used for a quick visualisation of characteristics of new events detected by EWS.

GAIA

Gaia is the European Space Agency's corner-stone mission, aiming mainly at all-sky high-precision astrometry of stars down to V=20 mag. It will also collect spectrometry and photometry data of billions of stars. Over its five years of operation, Gaia will scan the entire sky and will return to each place, on average, around 80 times. However, some places of the sky (around the nodes of the Gaia's spinning axes) will be observed up to 250 times.

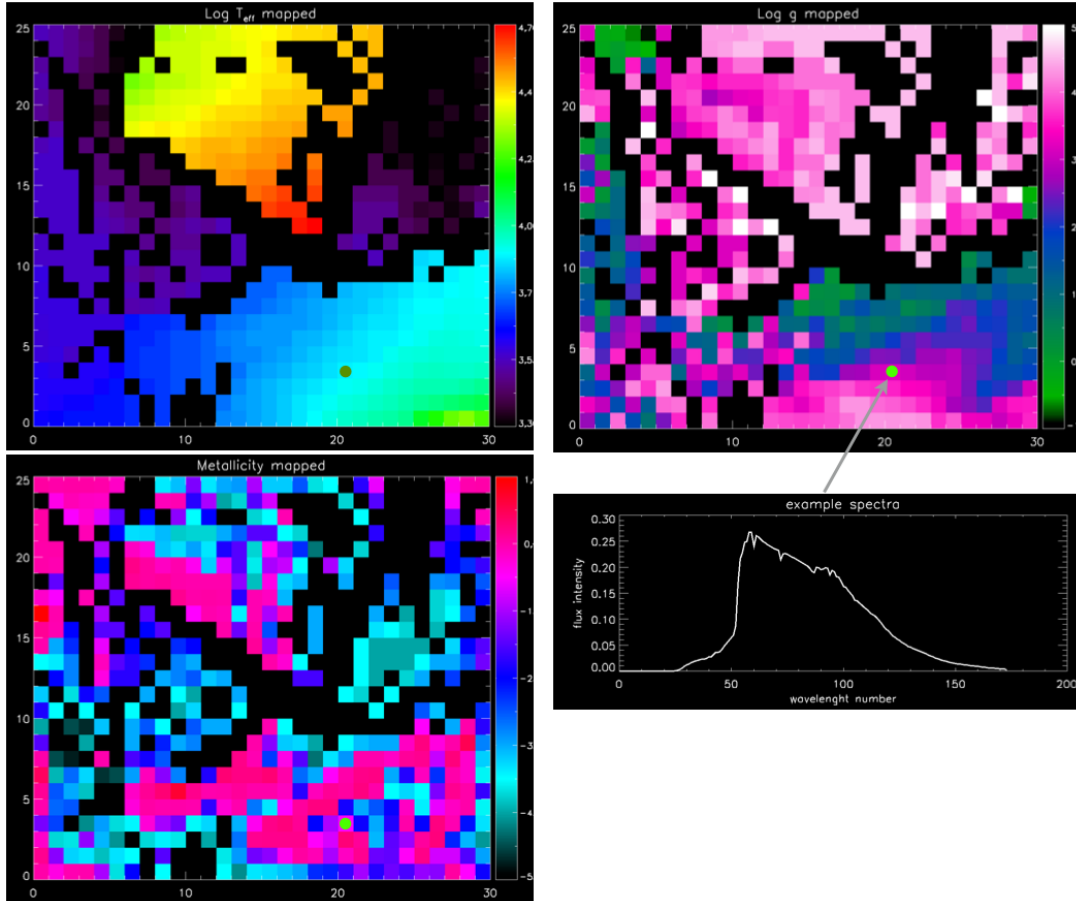


FIGURE 4. Maps of Basel library spectra physical parameters mapped back on the SOM. Any spectrum shown to the SOM can have its parameters derived immediately from such maps. Exemplary spectra and its position on the maps is shown with green dot.

Gaia Science Alerts

Photometric measurements collected by the satellite will be available for the first analysis after about 24 hours. Science Alerts are responsible for rapid detection of flux anomalies in the initial data caused by, e.g. supernovae, dwarf novae or microlensing events. Science Alerts tools will also analyse the accumulated data and will use SOMs to detect changes and anomalies in the spectra of the sources due to, for example, variability in eclipsing binaries or pulsating stars. Additionally, SOM will immediately allow for detection of new kind of spectra, not similar to any of known spectra [5].

The SOM was trained with about 8000 spectra from the Basel library ([6]) covering wide ranges of temperature, surface gravity and metallicity. These parameters were then mapped back on the map (see Figure 4). Multidimensional sorting ability of the SOM can be easily seen. With a SOM trained, for any spectrum a winning node can be found (i.e. the most similar spectrum) and its temperature, surface gravity and metallicity can be simply read out from the maps. Gaia Science Alerts will be tracking changes of the physical parameters of sources and will alert on any anomalous behaviour. Also shown an exemplary spectrum with known $\log T_{\text{eff}}=3.9$, $\log g=3.5$ and $\text{Fe}/\text{H}=-0.15$, which was identified to be the most similar to the node (20,3) of the SOM. Its position on parameters' maps is marked with green dots. It corresponds exactly to the known parameters of the spectrum. This useful feature of SOMs can have wide and numerous applications in rapid classification of spectra.

APPLICATIONS OF SOMS ON LIGHT CURVES

Future astronomical surveys will be observing billions of stars. The light curves of millions of variable stars sharing a similar shape (e.g. RR Lyrae, eclipsing binaries) will be observed at different phases. SOM can be used for completing the missing parts of the light curve when the sampling is not frequent enough but the number of available light curves is large (like in the case of Gaia's photometric data stream). This SOM was trained with two different RR-Lyra-like simulated variables observed sparsely at random phases (big dots on Figure 5). After training with several hundreds of patterns, such SOM is now capable of distinguishing between the two types of variables and can fill the gaps between the observed data points.

Another SOM shown in Figure 6 was trained with patterns comprising of light curves of three different kinds of simulated variable stars (upper right panel). As in real world, every light curve was shifted to start at a random phase. To deal with such offsets, this SOM first uses a cross-correlation function to match the phase of the pattern and only then uses the Euclidian distance to measure similarity. This novelty approach is important in training SOMs with light curves and has a great potential in classifying variable stars in real data sets.

The SOM trained to recognise three types of variable light curves (above) is also capable of figuring out the class of a new data pattern using only a few first data points. Grey shaded points in Figure 6 show changes in the classification as more data points are being added to the input pattern. Each incomplete pattern was shown to the map and the winner was found. The track (in red on the map) converged to the correct answer

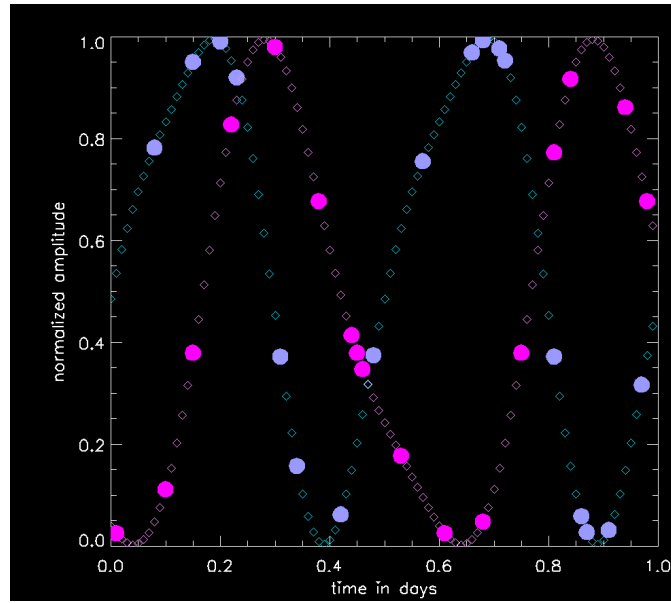


FIGURE 5. SOM can fill the gaps (empty points) between sparsely sampled data (filled points).

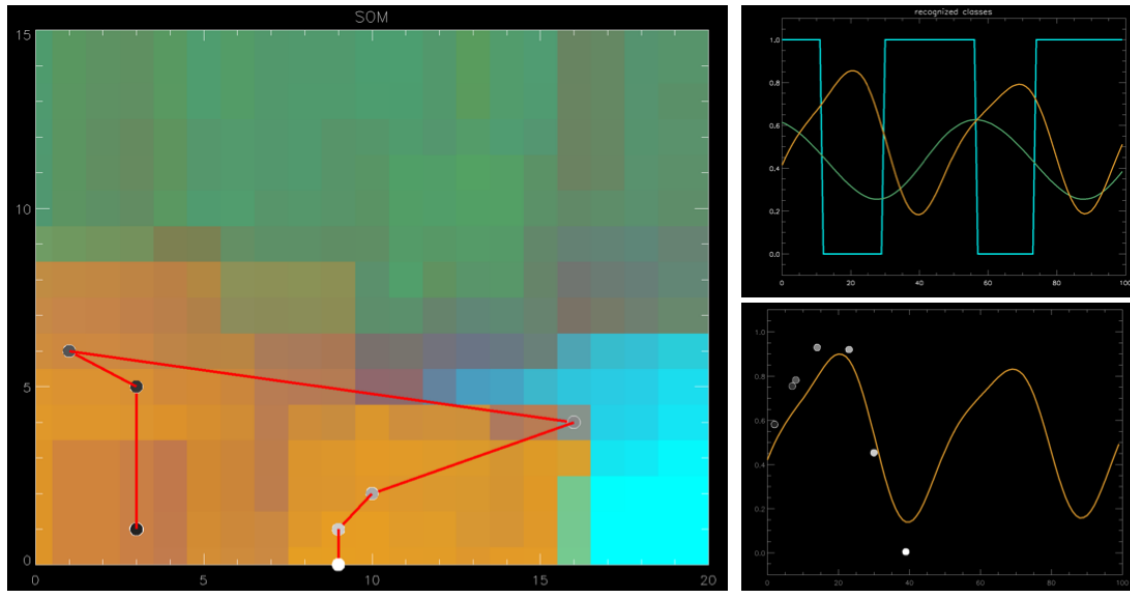


FIGURE 6. SOM trained on three different type of variability (upper right panel) with a track showing convergence of recognition of incomplete data. Black-grey-white dots are consequent data points added to the light curve(lower right panel).

(node 9,0) already when only 7 points of the light curve were present. This SOM feature can be applied for real-time classification of variable stars in incoming data.

REFERENCES

1. T. Kohonen, *Biological Cybernetics* **43**, 59 (1982).
2. B. Paczyński, *ARAA* **34**, 419–460 (1996), arXiv:astro-ph/9604011.
3. A. Udalski, *Acta Astronomica* **53**, 291–305 (2003), arXiv:astro-ph/0401123.
4. B. S. Gaudi, D. P. Bennett, A. Udalski, A. Gould, and et al., *Science* **319**, 927 (2008), arXiv:0802.1920.
5. N. W. Evans, and V. Belokurov, “A Prototype for Science Alerts,” in *The Three Dimensional Universe with GAIA*, eds. M. Perryman, C. Turon, ESA SP-576, 2004, arXiv:astro-ph/0411439.
6. E. Lastennet, T. Lejeune, E. Oblak, P. Westera, and R. Buser, *Astrophysics and Space Science* **280**, 83–87 (2002), arXiv:astro-ph/0112461.